$$E_{ij} = \begin{cases} \infty, & \text{for } r_{ij} < 3 \\ E^{rep}, & \text{for } 3 \le r_{ij} < R_{ij}^{rep} \\ \varepsilon_{ij}, & \text{for } R_{ij}^{rep} \le r_{ig} < R_{ig} \\ 0. & \text{for } R_{ig} < r_{ig} \end{cases} \quad (8)$$

where ε_{ij} were the pairwise interaction parameters, r_{ig} was the distance between chain beads i and j, E^{rep} =3kT was a constant repulsive term operating at very short distances, and R_{ij}^{rep} and R_{ig} were the cut-off values that depend on amino acid type. The values of these cut-off parameters were provided in Table VII.

Table VII. Compilation of pairwise cut-off distances for pairwise interactions

A_i	A_j	R_{ij}^{rep} (Å)	R_{ig} (Å)
Small[b]	Small	4.35[b]	5.97
Large[c]	Large	4.83	6.80
Other	Combinations[d]	4.57	6.32

[a] Small amino acids are: Gly, Ala, Ser, Cys.

[b] This value corresponds to the excluded volume radius of three lattice units; therefore, for pairs of small amino acids, the soft-core envelope does not exist.

[c] Large amino acids are Phe, Tyr, Trp.

[d] Small-large, other (than small or large)-large, other-small.

The interaction parameters depended not only on amino acid identity, but also on their positions in the polypeptide chain because the derivation of the potentials also used evolutionary information. A more detailed description of the derivation of these potentials is found elsewhere.[18] The total energy contribution from the pairwise interactions was therefore calculated as follows:

$$E_{pair} = \Sigma \Sigma E_{ij} \quad (9)$$

where the summations were over all j>i pairs of residues.

5. Multibody potentials

The hydrophobic interactions in this model were partially accounted for by pairwise interactions between residues; however, this was not sufficient to generate well packed proteins. Thus, a surface exposure based statistical potential was developed according to the following scheme: Each model residue was assigned 24 surface contact points. A specific subset of these contact points became occupied upon contact with other residues. The main chain Cα atoms contributed separately to the coverage of a given residue. The positions of the Cα atom could be quite well approximated given the positions of three consecutive side chain beads.[17] Some contact points could be multiply occupied. The fraction of non-occupied surface points defined the exposed fraction of a given side chain. Potentials could be derived from a statistical analysis of the protein structures for which the solvent exposure had been determined on the atomic level. The total surface energy was computed as follows:

$$E_{surface} = \Sigma E_b(A_i, a_i) \qquad (10)$$

where a_i was the covered fraction of the residue A_i and $E_b(A_i, a_i)$ was the statistical potential when amino acid type A had a_i of its surface points occupied, *i.e.*, the covered fraction of its surface was equal to $a_i/24$.

Studying the distribution of inter-residue contacts in globular proteins, various amino acids have been found to have different tendencies to pack in a parallel or antiparallel fashion. A contact between residues i and j was considered to be "parallel" when $(\mathbf{v}_{i-1} - \mathbf{v}_i)\bullet(\mathbf{v}_{j-1} - \mathbf{v}_j)>0$, and "antiparallel" otherwise. Moreover, for a given residue there were strong correlations between the number of parallel and antiparallel contacts given the total number of contacts. Due to the reduced character of this model, the other contributions to the force field did not properly account for such effects. Therefore, the model force field was supplemented by the following multibody potential:

$$E_{multi} = \Sigma E_m(A,n_p,n_a) \quad (11)$$

where $E_m(A,n_p,n_a)$ was the value of the statistical potential for residue type A having n_p parallel and n_a antiparallel contacts. The reference state was a random distribution of contacts. The values along particular diagonals ($n_p + n_a = n_c$) were normalized such that the lowest energy for a diagonal was exactly equal to the value of statistical potentials derived from the distribution of the total number of contacts n_c for a given type of residue.

6. Total intrinsic conformational energy

The total internal conformational energy of the model chain was equal to:

$$E_{total} = E_{stiff}+E_{map}+0.875E_{H\text{-}bond}+0.75E_{short}+1.25E_{pair}+0.5E_{surface}+0.5E_{multi} \quad (12)$$

with the value of generic parameter $\varepsilon_{gen} = 1$ kT.

The relative scaling of various potentials was adjusted by a trial and error method in *ab initio* folding experiments performed for a few selected small proteins, 1 fna, the B domain of protein A and the B1 domain of protein G. The objective was to maintain low secondary structure content in the random coiled state and dense packing with a proper level of secondary structure in the collapsed globular state. For instance, the small 56-residue α/β protein G domain folded *ab initio* in about 30% of simulated annealing Monte Carlo simulations to a native-like structure with an RMSD from native in the range of 4 Å. The majority of the remaining misfolded conformations had native-like secondary structures, but they had topological errors, usually involving the wrong order of β-strands in the four-member β-sheet. The model is not sensitive to small variations in these scaling parameters.

Building the starting lattice model

A separate algorithm was used to build an initial lattice model from a given target sequence alignment to a template structure. Such alignments contain gaps and